June 2020

Surfacing Biased Portions of Multimedia Content Using Machine Learning

Daniel Keysers

Victor Carbune

Surfacing Biased Portions of Multimedia Content Using Machine Learning

ABSTRACT

Multimedia enables content creators to communicate information by adding nuance which is difficult to convey through written language via voice tone, camera angle, content highlighting, etc. However, it can be difficult for content consumers to discern biased opinions included within multimedia content. This disclosure describes techniques to automatically detect and surface such biased opinions within multimedia content. The process involves examining publicly available multimedia and/or text content related to a given piece of multimedia content to identify and flag biased portions. The identified biased portions are surfaced to the user via a suitable user interface mechanism.

KEYWORDS

- Multimedia content
- Biased content
- Bias detection
- Bias flagging
- Content recommendation
- Event coverage

BACKGROUND

Multimedia content, such as podcasts and videos, is a popular mechanism for information consumption. Compared to text alone, multimedia enables a content creator to communicate information by adding nuance, by varying the tone of voice, recording relevant snippets of public events, presenting a setting from different visual angles, etc. As a result, multimedia information

delivery can enrich the content consumer's understanding of the information being communicated.

At the same time, it can be difficult for content consumers to discern intentionally biased opinions included within such content at the time of consumption. Some examples of content that included biased aspects include:

- **Online videos that include product information**: A general discussion of the value of a generic product may lead to subsequent presentation of a specific branded product within the natural flow of a video clip, leaving the viewer wondering whether the content was an unbiased opinion or paid promotion of the product.
- **Shows that involve debates among contestants**: The questions, clarifications, and interventions of a debate moderator influence debate flow and can raise questions about whether the moderator acts intentionally to favor a specific debate participant.
- **Coverages of events**: Descriptions of events covered in podcasts or videos can be biased when portraying the event based on various relevant choices such as selection of individuals that are interviewed, non-verbal cues that are employed, emotional words used, images shown during the presentation, etc.

DESCRIPTION

This disclosure describes techniques to automatically detect and surface biased opinions included within multimedia content. The process involves examining publicly available multimedia and/or text content related to the given piece of multimedia content to flag potentially biased portions within it. Such potentially biased portions are conveyed via a suitable user interface mechanism, such as annotations, tooltips, pointers to relevant unbiased information resources, etc. that are presented before, during, or after content consumption.

The multimedia content under review is analyzed to extract metadata, such as detected objects, speech, themes and topics within the content, etc. For example, the analysis can be performed via suitable trained machine learning models, such as convolutional neural networks (CNN), recurrent neural networks (RNN), transformer neural networks, etc. Subsequently, publicly accessible materials that are related to the given multimedia content are identified and retrieved, e.g., by employing techniques such as recommender systems or collaborative filtering, that are used for recommendation of content related to a given piece of content.

The retrieved related material and the multimedia content under examination are analyzed using a trained machine learning classifier. Additionally, the classifier can be provided with relevant embeddings (generated from source audio/video using pretrained models) pertaining to the content corpus, such as text produced by speech recognition, tags produced by image analysis, high-level annotations that classify the content into classes of events, etc. The output of the classifier indicates whether a particular portion within the content that is being examined is likely to be biased. The portion(s) flagged for bias can be based on audio, video, or text within the multimedia content that is being inspected.

Any biased portions detected within the given multimedia content are surfaced to the user within the UI of content consumption, e.g., an audio or video player application, a multimedia viewer application, etc. Along with indicating that certain elements within the content are likely to be biased, the user can be shown relevant unbiased information that helps counter the bias. Such unbiased information can include pointers to relevant related materials.

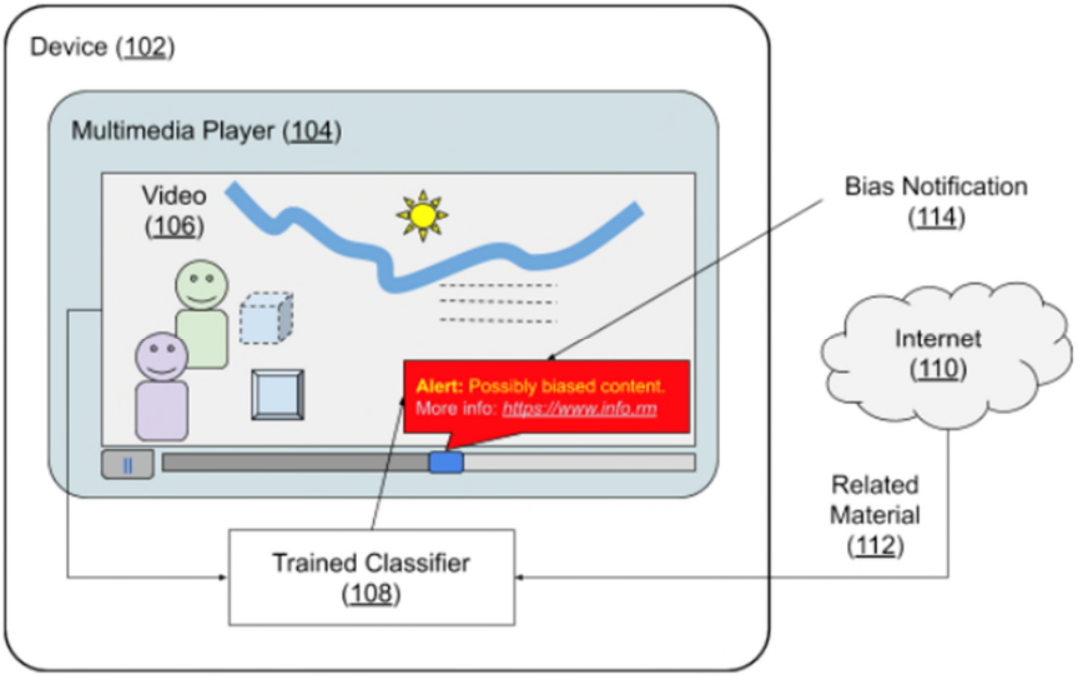

Fig. 1: Flagging potentially biased portions within multimedia content

Fig. 1 shows an example operational implementation of the techniques described in this disclosure. A user is viewing a video (106) via a multimedia player application (104) on a user device (102). The video content is examined via a trained machine learning based classifier (108) along with publicly accessible material related to the video (112) accessed via the Internet (110). While Fig. 1 shows the trained classifier as running on the user device, when appropriate permissions are obtained from the user, the classifier can run on a server, or a combination of the user device and the server. When the classifier detects that a portion within the video includes potentially biased information, the user is alerted via a bias notification (114). The bias notification can include pointers to relevant unbiased information.

The described techniques can be applied to the entire multimedia content as a whole, thus permitting bidirectional approaches conditional on information presented until the end of the

content. Such an operation can cover situations that are likely to escape user attention, such as advertising via subtle product placement.

The classifier can be trained via manually labeled data. For example, users can contribute to manual labeling by providing inputs that flag biased material during or immediately after content consumption. For example, during consumption, users can be presented questions, such as “Do you believe this material might be biased?” If the user permits, such queries can be triggered based on whether the user consumed content related to the content currently being consumed.

The techniques described in this disclosure can be incorporated within any application or platform that is used to serve and/or view multimedia content such as online video websites, podcast applications, etc. The techniques can also be integrated within online search engines, especially when the search results that are relevant to the user’s query include multimedia content. In such cases, the search results can be augmented to identify potential bias in specific search results.

Further to the descriptions above, a user may be provided with controls allowing the user to make an election as to both if and when systems, programs or features described herein may enable collection of user information (e.g., information about a user’s content viewing history, content that a user is currently viewing, a user’s preferences, etc.), and if the user is sent content or communications from a server. In addition, certain data may be treated in one or more ways before it is stored or used, so that personally identifiable information is removed. For example, a user’s identity may be treated so that no personally identifiable information can be determined for the user. Thus, the user may have control over what information is collected about the user, how that information is used, and what information is provided to the user.

CONCLUSION

This disclosure describes techniques to automatically detect and surface such biased opinions within multimedia content. The process involves examining publicly available multimedia and/or text content related to a given piece of multimedia content to identify and flag biased portions. The related material and the multimedia content under examination are analyzed using a trained machine learning classifier. The output of the classifier indicates whether a particular portion within the content being examined is likely to be biased. The identified biased portions are surfaced to the user via a suitable user interface mechanism.